چالش های تحلیل متن فارسی:

>> پیش پردازش صحیح. نرمالایز کردن

>> اینکه در چه سطحی توکن کنیم خیلی مهمه. توکن ها نه زیاد باید طویل باشن نه خیلی به سوی اتومیسیتی پیش برن. مثلا به این جمله توجه کنید:

دکتر حسن روحانی به نیویورک سفر می کند.

در این جمله، اگر توکنایز بتواند "دکتر حسن روحانی " را دتکت کند برای مدل سازی بسیار خوب است. ولی اگر هر کدام ازکلمات را یک توکن در نظر بگیرد کمک چندانی نمیکند.

>> چالش دیگر چالش تشخیص افعال ترکیبی است. مثلا : سفر می کند. غذا می خورد.

راه حل: روکیرد فازی

چالش های tf-idf

چالش 1:

بهتره قبل از دادن دیتا به این الگوریتم، با روش خودمون توکنایز کنیم. جمله زیر رو در نظر بگیرید:

دکتر روحانی به نیویورک سفر می کند.

اگر توکن کردن ما از قبل به این شکل بوده باشه:

"دکتر روحانی"، "به نیویورک"، "سفر می کند"

خیلی بهتر میتونه عبارات رو اسکور دهی کنه. مثلا میگرده دنبال تکرار ترم "دکتر روحانی"

راه حل:

من ابتدا کل ستون کلمات رو برداشتم. خوبی این دیتا ست اینه که توکنایز کردنش در سطح عبارت. مثلا همون دکتر روحانی. بعدش به عنوان لیست خوندمش. لیستی از کلمات. خوب حالا داکیومنت و جملات رو از توش استخراج کردم. هر جمله یک سند. چطوری؟ توی اون ستون گشتم دنبال punctutation ها. گفتم هرجا از این پانک ها دیدی ینی جمله تموم شده.

چالش 2:

تو این مساله خاص که ما یه دیتای لیبل شده داریم و میخوایم ازش ویژگی استخراج کنیم، آیا بعد از محاسله tf\_idf باید امتیاز هر ترم رو بزاریم تو همون دیتا ست اصلی برای طبقه بندی؟ دیتای تست تکلیفش چی میشه

چالش 3: ایست واژه ها باید تو این فرآیند باشند یا خیر؟

چرا co-occurrence برای استخراج مفاهیم مناسبه؟

به سند زیر توجه کنید.

روحانی به نیویورک سفر کرد. در نیویورک چندین روحانی زندگی می کند.

در این سند، تگ های مختلفی از توکن روحانی داریم. مثلا در جمله اول روحانی یک شخص است، این در حالی است که در جمله دوم روحانی یک حرفه. چطور می توان این دو توکن را تگ گزاری کرد؟ با روش هم رخدادی.

1. در این مرحله باید اندازه پنجره را مشخص کنیم. اندازه پنجره را برای هر توکن 2 در نظر می گیریم. مثلا برای توکن روحانی داریم:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| روحانی | به | نیویورک | سفر | کرد | در | نیویورک | چندین | روحانی | زندگی | می کند |
| روحانی | به | نیویورک | سفر | کرد | در | نیویورک | چندین | روحانی | زندگی | می کند |

1. سپس ماتریس هم رخدادی را می سازیم:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | روحانی | به | نیویورک | سفر | کرد | در | چندین | زندگی | می کند |
| روحانی | 0 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 1 |
| به | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| نیویورک | 2 | 1 | 0 | 1 | 2 | 1 | 1 | 0 | 0 |
| سفر | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| کرد | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 |
| در | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 |
| چندین | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| زندگی | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| می کند | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

1. حالا برای طبقه بندی باید زمانی که توکن روحانی را برای آموزش به مدل می دهیم، 2 توکن قبل و بعدش را نیز به مدل بدهیم. یعنی برداری به شکل زیر داریم:

|  |  |  |
| --- | --- | --- |
| روحانی | به | نیویورک |
| 012000111 | 101100000 | 210121100 |
| 012000111101100000210121100 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| نیویورک | چندین | روحانی | زندگی | می کند |
| 210121100 | 101001010 | 012000111 | 100000101 | 000000010 |
|  |  |  |  |  |

گاهی اوقات توکن هدف ما در ابتدای سند است ولی تکرار همان توکن در ابتدای سند نیز یافت می شود. با روش window اگر توکن ابتدای سند را در نظر بگیریم سمت راست را نمی توان کد گزاری کرد که از روش padding استفاده می کنیم. در این روش، باید اطلاعات به ظاهر به درد نخور استفاده شود. مثلا صفر که تاثیری در نتیجه ندارد. اگر داده رندم وارد کنیم ممکن است در خروجی تاثیر گزار باشد.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | روحانی | به | نیویورک | سفر | کرد | در | نیویورک | چندین | روحانی | زندگی | می کند |
|  |  | روحانی | به | نیویورک | سفر | کرد | در | نیویورک | چندین | روحانی | زندگی | می کند |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pad1 | Pad0 | روحانی | به | نیویورک |
| 000000000 | 000000000 | 012000111 | 101100000 | 210121100 |
| 000000000000000000012000111101100000210121100 | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| نیویورک | چندین | روحانی | زندگی | می کند |
| 210121100 | 101001010 | 012000111 | 100000101 | 000000010 |
| 210121100101001010012000111100000101000000010 | | | | |

چالش های ماتریس هم رخدادی:

1. اگر کانتکست را کل مخزن متن در نظر بگیریم، ابعاد مجموعه داده بسیار بالا می شود. مثلا در صورتی که 700 رکورد منحصر بفرد داشته باشیم، این عدد ضرب در اندازه پنجره برای پیمایش کانتکس می شود که مقداری برابر با 3500 بعد به دیتا ست می دهد.
2. تجزیه مخزن متن به k زیر سند یا کانتکست. در این روش ما به اندازه دلخواه اندازه سند را تغییر می دهیم. مشکلی که در این روش وجود دارد برابر نبودن توکن های موجود در هر سند می باشد. که موجب می شود ابعاد کل توکن ها در مجموعه داده برابر نباشد. راه حل این است که تعداد توکن های متن را به گونه ای تهیه کنیم تا هر کانتکست تعداد توکن های برابری داشته باشد. مثلا متن با 1000 توکن به 10 توکن 100 تایی تقسیم می شود.
3. اگر متن به 10 قسمت 100 تایی تقسیم شود، ممکن است در مرز این فولد ها توکن های معنا دار شبیه به هم قرار داشته باشن که با تجزیه کردن متن به این فولد ها این توکن ها از هم جدا می شوند و در کانتکست خودشان بررسی می شوند.